Open Refine Tutorial 2 of 3:
Cleaning Numerical Data

In this tutorial on cleaning messy data we'll see how Open Refine can help catch and resolve numerical mistakes.

We'll use the same dataset used in the first tutorial on Open Refine. The first tutorial also has instructions on how to download and install Open Refine.

To access the dataset go here: https://data.lacounty.gov/Arts-and-Culture/Los-Angeles-County-Civic-Art-Collection/9qv9-dayy

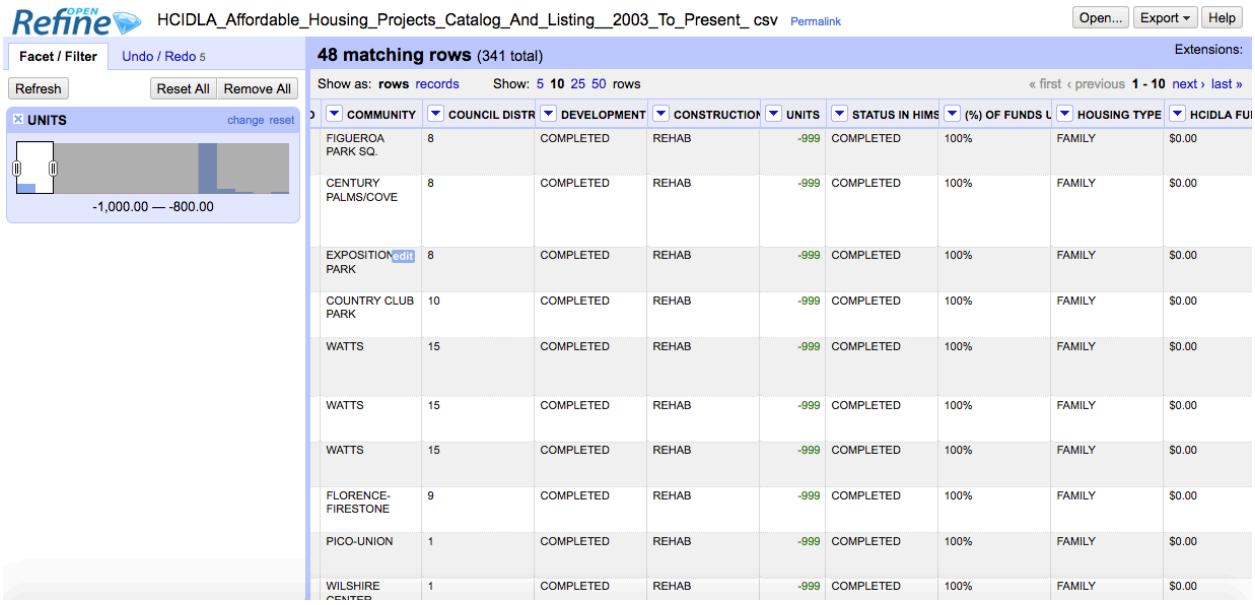
Numeric Faceting

Numeric faceting is one way to detect and fix numerical errors in the dataset. Begin by examining the 'Units' column of the Affordable Housing Projects Catalog dataset. Go to 'Facet' → 'Numeric Facet'.

Note: if the facet returns zero results, that means the column can't read these cells as text. To correct this, go to the column heading's drop down menu, and go to 'Edit cells' → 'Common Transforms' → 'To Number'.

Numeric and Timeline facets display histograms instead of lists of values. The histograms include 'drag and drop' controls you can use to set a start and end range to filter the data displayed.

You can see that there are, oddly, negative numbers in the results of the values in the 'Units' column. To isolate the cells that have negative numbers, toggle the histogram sliders so that they bracket the results at the far left of the chart – the numbers less than zero.



We can see that there are 48 rows with -999 living units, clearly an error. Because we don't know the source of this error, we aren't able to clean it with the correct information. (At best you could make a note on any analysis regarding 'Units' that qualifies the set as incomplete.)

Let's say we want to remove the results with negative numbers from our analysis. After isolating the facet to rows with negative numbers, go to the drop down menu in the 'All' column. Click on 'Edit rows' → 'Remove all matching rows'. This will delete all the rows that have -999 in the 'units' column.

Converting to Numbers

Another common problem with numerical data is that they are often listed as text descriptions of numeric values (i.e. $500 million or 500,000,000) rather than pure numeric values (i.e. 500000000) that can be subject to analysis.

In this spreadsheet, all economic data contains dollar signs, so Open Refine reads it as text. Transforming the cells to read as numbers will not work so long as the dollar sign is place.

Luckily, Open Refine has a 'replace' function that can eliminate any unwanted text. Start with the 'TDC' column – the column containing the total cost of each housing construction. Click **Edit cells → Transform.**

The pop up window that displays allows you to now type in a command that can replace sequences of characters:

```
value.replace("$","")
```

This command replaces the dollar sign with nothing and so eliminates it from each cell. You can see what the transformation will look like in the preview below the 'Expression' text box. Now click 'OK'.

Now you'll need to convert the column to read as numbers. To perform this action go to the column heading's drop down menu, then to 'Edit cells' → 'Common Transforms' → 'To Number'. Now you're ready to make numeric facets from this column.

Note: The replace command is an example of a 'Transformation', which allows you to manipulate data in a column. Transformations typically use a programming language called Google Refine Expression Language (GREL). Several advanced features like this are available. If you want to explore more commands visit the Google Refine Expression Language (GREL) reference [https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users#reference].

Credits:

Written by Morgan Currie, September 2016.